BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 9: e183132
doi: 10.3897/biss.9.183132

OPEN ACCESS

Conference Abstract

MfN DataHub – a Centralized Service for Automated Biodiversity Data Integration at the Museum for Natural History Berlin

Majid Vafadar ‡

‡ Museum für Naturkunde Berlin, Berlin, Germany

Corresponding author: Majid Vafadar (majid.vafadar@mfn.berlin)

Received: 20 Dec 2025 | Published: 23 Dec 2025

Citation: Vafadar M (2025) MfN DataHub – a Centralized Service for Automated Biodiversity Data Integration at the Museum for Natural History Berlin. Biodiversity Information Science and Standards 9: e183132. https://doi.org/10.3897/biss.9.183132

Abstract

The Museum für Naturkunde Berlin (MfN) DataHub*[1] is an open-source web service and workflow engine developed to execute automated data-integration and migration workflows in continuous and parallel scenarios. Data migration and integration remain major challenges in publishing biodiversity data that follow international standards. To overcome these, a centralized service was created to coordinate and concentrate the computational power required for large-scale data transformation. Deployed at the Museum für Naturkunde Berlin, it now serves as the core of the institution's scientific data-management infrastructure.

This service supports digitization pipelines by iterating through datasets and files to integrate them into designated target systems. Following the ETL (Extract, Transform, Load) (Moreau 2015) principle, it extracts data from internal databases and shared storages via secure protocols such as SMB*[2] and SFTP*[3], transforms and validates them, and loads the compliant outputs through target-system API endpoints. The overall architecture and data flow are shown in Fig. 1.

Operations are controlled through a web dashboard that allows execution and monitoring of pipelines either manually, automatically, or with AI-agent assistance. Implemented using the Django Web Framework, the service runs modular Python scripts and exposes

© Vafadar M. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

all functions through RESTful APIs. A MCP*[4] server provides an AI-readable interface, enabling both human- and machine-driven operations.

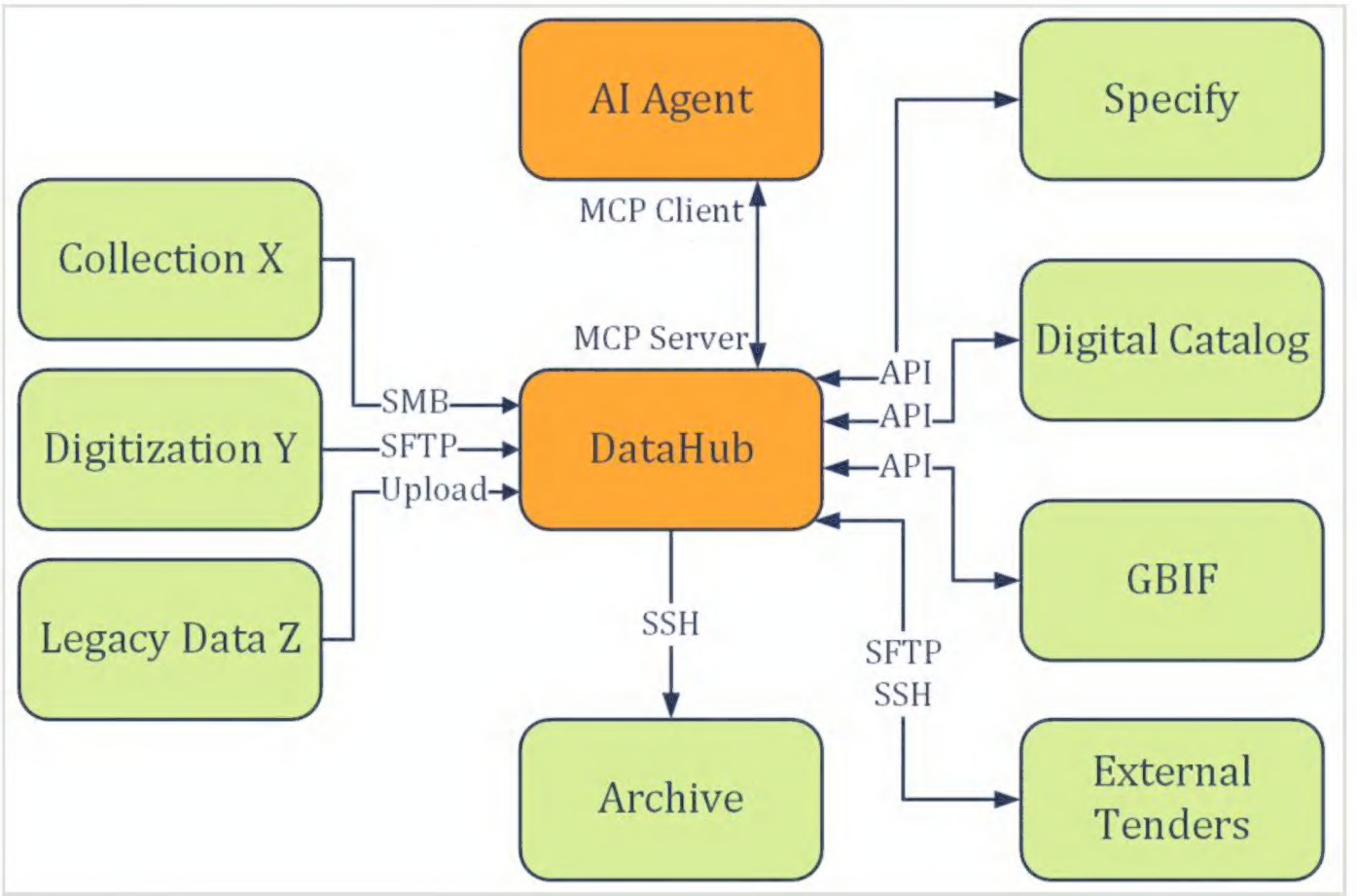

Figure 1.

Overview of the MfN DataHub architecture and data flows: Data sources connect to the DataHub through SMB*[2] or SFTP*[3]. Target systems like Specify Collection Management System, the museum's Digital Catalog, GBIF (Global Biodiversity Information Facility) interact via API or SSH*[5]. An AI Agent communicates through the MCP*[4] server for workflow orchestration. From an original diagram made by M. Vafadar for the Museum für Naturkunde Berlin (MfN), and adapted for use here. Licensed under CC BY 4.0.

Connected to the museum's storage systems, the DataHub validates, enriches, and transforms data with a dedicated validator ensuring each record meets predefined structural and semantic rules. A persistent integration pipeline imports datasets into the museum's Specify collection management system and its digital catalog, making data accessible for research and public use. It also prepares standardized packages for external partners such as GBIF (Global Biodiversity Information Facility), following the Darwin Core (Wieczorek et al. 2012) format and specific project requirements. Integration with field-data applications like ODK (Open Data Kit) ensures mobile data collection can enter the same pipeline. All operational steps, including validation, transformation, and API transactions are fully logged for transparency and reproducibility of the operation.

A key innovation is the AI-integration layer, linking through the MCP*[4] server to an AI agent built with LangChain library and Qwen3 LLM*[6] model, executed locally via Ollama platform. This component assists with workflow orchestration by optimizing tasks order, resource allocation, error recovery and live reports thereby reducing manual supervision.

By combining ETL*[7] pipelines with AI-assisted orchestration, the DataHub provides a flexible, scalable engine adaptable to different collection domains. Its modular and open-source design promotes reproducibility and extension to new systems. The centralized architecture enhances data quality and FAIR (Findability, Accessibility, Interoperability, and Reusability) Wilkinson et al. 2016 compliance, offering a practical and scalable solution that can be deployed in natural-history institutions aiming to modernize their digital-collection infrastructures and ensure the continuous availability of reliable, high-quality biodiversity data. The open-source code and technical documentation of the MfN DataHub are available on the museum's GitHub repository.*[1]

Keywords

mfn datahub, data migration, ETL, biodiversity informatics, natural history collections, digitization pipelines, workflow automation, data validation, artificial intelligence, MCP, LangChain, Qwen3, Darwin Core, GBIF, FAIR data, open source

Presenting author

Majid Vafadar

Presented at

Living Data 2025

Acknowledgements

The author would like to thank the Museum für Naturkunde Berlin (MfN) for supporting the development and deployment of the DataHub as part of the museum's digital infrastructure initiatives. Special thanks go to the Digitization and IT teams at the MfN for their collaboration in testing and integrating the system with existing workflows. The presentation of this work at Living Data 2025 Conference in Bogotá, Colombia, was made possible through the institutional support of MfN's Data Management department.

Hosting institution

Museum für Naturkunde Berlin

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Moreau L (2015) The W3C PROV Family of Specifications. World Wide Web Consortium (W3C)
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An evolving community-developed biodiversity data standard. PLoS ONE 7 (1): 29715. https://doi.org/10.1371/journal.pone.0029715
- Wilkinson MD, Dumontier M, Aalbersberg IJ (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18

Endnotes

*1 https://github.com/MfN-Berlin/mfn-datahub-public
*2 Server Message Block
*3 Secure File Transfer Protocol
*4 Model Context Protocol
*5 Secure Shell
*6 Large Language Model
*7 Extract, Transform, Load